Wie stabil sind Umfragedaten? Beschreibung und erste Ergebnisse der Test-Retest-Studie zum ALLBUS 1984

Porst, Rolf; Zeifang, Klaus; Koch, Achim

Wie stabil sind Umfragedaten? Beschreibung und erste Ergebnisse der Test-Retest-Studie zum ALLBUS 1984

Grundlagenforschung im Bereich von Umfragemethodologie und Umfragemethodik ist eine von mehreren zentralen Aufgaben des Forschungsprogramms ALLBUS (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften). Mit jeder ALLBUS-Umfrage ist von daher eine "begleitende Methodenstudie" verbunden gewesen, mit deren Ergebnissen die Diskussion zentraler Fragestellungen der Umfrageforschung empirisch vorangetrieben werden soll. Begleitende Methodenstudie zum ALLBUS 1984 war die sog. "Test-Retest-Studie" gewesen, bei der eine Teilstichprobe der Befragten des ALLBUS 1984 an zwei Nachbefragungen teilgenommen hat. Ziel der Studie war die Ermittlung der Stabilität von Umfragedaten. Die Primärauswertung erfolgte durch eine Arbeitsgruppe, die eigens zu diesem Zweck gebildet worden war. Die Ergebnisse ihrer Arbeit werden in einem Sonderband der Zeitschrift "Sociological Methods and Research" veröffentlicht (Heft 2/1987). Die Daten der Test-Retest-Studie zum ALLBUS 1984 werden in Kürze beim Zentralarchiv für empirische Sozialforschung an der Universität zu Köln archiviert und sind dann für Interessenten frei zugänglich.

Im folgenden werden zunächst die Konzeption und die Realisierung der Test-Retest-Studie beschrieben. Dieser eher konventionelle "Methodenbericht" endet mit einem systematischen Querschnittsvergleich zwischen denjenigen Befragten, die an der Hauptstudie zum ALLBUS 1984, nicht aber an den Nachbefragungen teilgenommen haben (N=2.850), und denjenigen Befragten, die später auch an beiden Nachbefragungen der Test-Retest-Studie teilgenommen haben (N=154). Geht es dabei um die Frage der Stichprobenqualität des Test-Retest-Samples, dreht sich der eigentlich substantielle Teil dieses Artikels um die Frage der Antwortstabilität über die Zeit: Wie stabil sind Umfragedaten?

1. Die Test-Retest-Studie zum ALLBUS 1984 - Konzeption, Realisierung und Qualität der Stichprobe

Alle Überlegungen zur Konzeption und Durchführung der Test-Retest-Studie sind ausgegangen vom Zeitreihencharakter der ALLBUS-Daten. Will man sich, insbesondere im Rahmen komplexer Analysemodelle, versichern, daß es sich bei gemessenen Veränderungen über die Zeit tatsächlich um Wandel und nicht nur um methodische Artefakte handelt, sind Angaben über die methodische Qualität der Meßinstrumente unabdingbar. Als Indikator für die Stabilität von Meßinstrumenten über die Zeit eignet sich die Test-Retest-Reliabilität.

1.1 Konzeption und Begründung der Test-Retest-Studie

Die Reliabilität von Meßinstrumenten gilt, neben Validität, als zentrale Voraussetzung für die Qualität der erhobenen Daten und damit als zentrale Voraussetzung für das Testen von Hypothesen. Die wichtigsten statistischen

Grundlagen zur Messung von Reliabilität wurden in der psychologischen Testtheorie entwickelt und auch von soziologischer Seite übernommen (Heise/Bohrnstedt 1970:104-129), doch werden Reliabilitäten in Arbeiten der empirischen Sozialforschung häufig nicht berechnet oder zumindest nicht publiziert (Porst/Schmidt 1982:9).

Das aus der klassischen Testtheorie (Lord/Novick 1968) bekannte Konzept der Reliabilität setzt sich a) mit der Stabilität eines Meßinstruments über die Zeit und/oder b) mit der internen Konsistenz eines Sets von Items auseinander, von denen anzunehmen ist, daß sie ein gemeinsames latentes Konstrukt messen. Reliabilität ist definiert als die quadrierte Korrelation zwischen gemessenen und "wahren" Werten von Variablen bzw. als Maß für das Verhältnis der Varianz der wahren Werte zur Varianz der beobachteten Werte. Mit anderen Worten: Ein hoher Reliabilitätskoeffizient ist ein Indikator für eine hohe Interkorrelation zwischen dem empirischen Wert einer Variablen und ihrem wahren Wert (Lord/Novick 1968).

Erfolgt die Reliabilitätsbestimmung auf der Basis zeitverschobener Messungen bei Konstanz der Befragungseinheiten (also im Panel), kann von Stabilitätsmessung gesprochen werden (vgl. Wegener 1983:54ff.): Stabilität beschreibt das Ausmaß, in dem Befragungspersonen eine bestimmte Frage über mehrere Erhebungszeitpunkte hinweg konsistent, also mit dem gleichen Response, beantworten.

Wenngleich die Bestimmung der Test-Retest-Reliabilität als der einfachste Fall der Stabilitätsmessung bezeichnet wird (Wegener 1983:54), erweist sie sich dennoch aus mehreren Gründen als schwierig:

1. Veränderungen gemessener Werte über die Zeit können Folge von Meßfehlern, aber auch Folge tatsächlichen Wandels sein, oder von beidem zusammen.
2. Veränderungen gemessener Werte können Folge unterschiedlicher kontextueller Bedingungen bei den Interviews der unterschiedlichen Befragungswellen sein (z.B. aufgrund der unterschiedlichen Anwesenheit weiterer Personen beim Interview).
3. Konstanz gemessener Werte kann dadurch entstehen, daß beim Befragten ein Lern- oder Erinnerungseffekt auftritt (der Befragte könnte sich z.B. bewußt bemühen, die gleichen Antworten zu geben wie beim ersten Interview, auch wenn sich seine tatsächliche Einstellung zu einem bestimmten Problem seit damals verändert hat).
4. Veränderungen gemessener Werte können einfach dadurch entstanden sein, daß sich der Befragte nach dem ersten Interview mit dem Gegenstand des Interviews intensiver befaßt hat und erst dadurch - also mithin als Folge der Erstbefragung - eine Meinungs- oder Einstellungsänderung aufgetreten ist (vgl. Campbell/Stanley 1966, Carmines/Zeller 1979).

Da im ALLBUS als einer auf Replikation basierenden Studie bestimmte Erhebungsinstrumente regelmäßig eingesetzt werden, erschien es unbedingt notwendig, Informationen über die Test-Retest-Reliabilität zumindest dieser Standardinstrumente zu erhalten. Neben diesem eher forschungspragmatischen Argument gibt es mindestens zwei systematische Argumente, die die Durchführung der Test-Retest-Studie im Zusammenhang mit dem ALLBUS als einer Datenbasis für Zeitreihenanalysen auf der Grundlage replikativer Querschnitte nicht nur sinnvoll, sondern unabdingbar erscheinen ließen.

So kann, erstens, mit Hilfe dieser Panel-Studie ermittelt werden, wie konsistent Befragungspersonen die gleichen Fragen beantworten, wenn sie innerhalb relativ kurzer Zeit mehrmals mit ihnen konfrontiert werden: Die Ergebnisse der Test-Retest-Studie ermöglichen Aussagen über kurzfristige Veränderungen oder über die Stabilität von Merkmalen und Einstellungen auf Individualebene. Unseres Wissens ist diese Fragestellung in einer allgemeinen Bevölkerungsumfrage bisher nicht untersucht worden.

Bei Anwendung von Strukturgleichungsmodellen mit latenten Variablen kann, zweitens, zwischen der Stabilität der "wahren" Werte (wahrem Wandel) und Stabilität der meßfehlerbehafteten beobachteten Werte (klassische Test-Retest-Reliabilität) unterschieden werden (vgl. Heise 1969, Wiley/Wiley 1970).

1.2 Design der Test-Retest-Studie

Die Diskussion um das Design der Test-Retest-Studie konzentrierte sich vor allem auf die Vorstellungen über das Feld, insbesondere über die Zahl der Erhebungen und die Zeitabstände zwischen den Erhebungen, den Stichprobenplan und die Stichprobengröße sowie das Erhebungsinstrument.

1.2.1 Vorstellungen über das Feld

In der einschlägigen Literatur besteht Übereinstimmung, daß bei Vorlage von nur einem Indikator pro Konstrukt mindestens drei Erhebungszeitpunkte vorliegen müssen, um die gleichzeitige Schätzung von Reliabilität und Stabilität der wahren Werte zu ermöglichen. (vgl. Heise 1969, Arminger 1976, Kessler/Greenberg 1981).

Da die Haupterhebung des ALLBUS 1984 als erste Welle des Panels angenommen wurde, waren für die Test-Retest-Studie demzufolge zwei Nachbefragungen erforderlich. Entscheidend war, in welchem zeitlichen Abstand die Nachbefragungen stattfinden sollten.

Die Abstände zwischen den Erhebungen eines Panels hängen sehr stark von den Zielsetzungen einer spezifischen Studie ab; allgemein gültige Aussagen können nicht getroffen werden. Die Regelung, die für die Test-Retest-Studie gefunden wurde, basierte sowohl auf theoretischen als auch auf pragmatischen Überlegungen. Die Abstände zwischen den Wellen sollten relativ kurz sein, um das Ausmaß tatsächlicher Veränderungen zu minimieren. Allerdings sollten die Zeitabstände auch nicht zu kurz sein, weil das Antwortverhalten sonst zu stark von Erinnerungs- oder Lerneffekten der Befragten überlagert sein könnte. Da die gesamte Feldzeit des ALLBUS 1984 ohnehin nicht unangemessen ausgedehnt werden konnte, wurde schließlich entschieden, daß jede Person der Stichprobe jeweils nach exakt vier Wochen zum erstenmal, nach exakt weiteren vier Wochen zum zweitenmal nachbefragt werden sollte, wobei eine geringe Varianz im Falle kurzfristigen Nicht-Erreichens einkalkuliert wurde.

1.2.2 Stichprobenplan und Stichprobengröße

Das Stichprobenverfahren sollte so angelegt sein, daß als Ergebnis der dritten Panel-Welle noch mindestens 150 vollständig realisierte Interviews vorliegen sollten. Diese Zahl, die später auch tatsächlich erreicht werden konnte, hat sich alles in allem - insbesondere für die Analyse von Subgruppen - immer noch als relativ niedrig erwiesen, konnte allerdings im Rahmen einer realistischen Kostenkalkulation für die Gesamtstudie nicht höher angesetzt werden.

Grundgesamtheit der zweiten Welle sollten zunächst alle Personen sein, für die in der Haupterhebung ein vollständiges Interview zustande gekommen sein würde und die sich zur Teilnahme an weiteren Befragungen bereit erklären würden. Grundgesamtheit der dritten Welle sollten alle Teilnehmer an der zweiten Welle sein.

Die so gefaßte Grundgesamtheit der Befragten für die zweite Welle mußte dann allerdings aus Kostengründen neu definiert werden. Nicht mehr alle Befragten, für die in der Hauptstudie ein vollständiges Interview realisiert wur-

de, sollten zur Grundgesamtheit gehören, sondern nur noch die Befragten eines der drei in der Hauptstudie eingesetzten Stichprobennetze.[1)] Ausgesucht wurde das Netz, in dem in der Hauptstudie acht Brutto-Adressen als Kontaktadressen bearbeitet werden sollten (in den beiden anderen Netzen sollten nur je 7 Adressen bearbeitet werden). Zum Einsatz sollten alle 210 sample points dieses Netzes kommen.

In jedem dieser 210 sample points sollte von allen Befragten der Haupterhebung die Bereitschaft zur Teilnahme an zwei Wiederholungsbefragungen erbeten werden. Je zwei der acht Adressen jedes der 210 sample points sollten für die Nacherhebungen ausgewählt werden. Die Interviewer sollten zum Zeitpunkt der Haupterhebung selbst nicht wissen, daß es dort definitiv zu Nachbefragungen kommen sollte.

Damit hätte sich ein Brutto-Ansatz von insgesamt 420 Kontaktadressen (210 sample points x 2 Adressen) als Ausgangspunkt für die Test-Retest-Studie ergeben. Bei einer erwarteten Ausschöpfung von 70% für die ALLBUS-Hauptstudie reduzierte sich die Anzahl der Adressen auf 294.

Von diesen 294 sollten sich - so die Schätzung - 235 oder 80% nach dem Interview in der Hauptstudie zur Teilnahme an weiteren Befragungen bereit erklären. Von diesen 235 sollten dann - wiederum geschätzt - ca. 85% tatsächlich in der zweiten Welle teilnehmen; die Zahl der realisierten Interviews nach der zweiten Welle wurde somit auf ca. 200 festgelegt. Wenn - so die weitere Schätzung - von diesen 200 wiederum 75% auch in der dritten Welle teilnähmen, wäre die angestrebte Stichprobengröße von 150 vollständigen Interviews in der dritten Welle realisiert.

1.3 Das Erhebungsinstrument

Erhebungsinstrument der ersten Welle des Panels war der reguläre Fragebogen zum ALLBUS 1984. Gemäß den datenschutzrechtlichen Bestimmungen enthielt er am Ende eine Erklärung über die Bereitschaft zur Teilnahme an weiteren Befragungen.

Das Instrument, mit dem die Nachbefragungen durchgeführt wurden, war eine auf etwa die Hälfte der Befragungszeit reduzierte Version des Fragebogens der Hauptstudie. Die Fragen, die in der Nachbefragung zum Einsatz kamen,

wurden unter verschiedenen Gesichtspunkten ausgewählt. In erster Linie wurden Fragen berücksichtigt, die als Standardinstrumente des ALLBUS-Programms gelten können, also Fragen, die bereits innerhalb von ALLBUS-Umfragen repliziert worden waren. Dabei sollten Fragen unterschiedlichen Skalenniveaus zum Einsatz kommen, und zwar sowohl Fragen aus den inhaltlichen Bereichen als auch demographische Fragen.

1.4 Realisierung der Studie

Die angestrebte Stichprobengröße von 150 vollständig realisierten Interviews in der dritten Welle konnte, trotz einer nicht ganz den Erwartungen entsprechenden Teilnahmebereitschaft der Befragten nach der Hauptstudie, letztlich doch realisiert werden. Statt der ursprünglich 235 Personen erklärten sich nach dem ersten Interview nur 210 zur Teilnahme an weiteren Befragungen bereit. Von diesen 210 Personen konnten in der zweiten Welle 181 (oder 86%) befragt werden, in der dritten Welle noch einmal 154 (oder 85% von 181). Anders ausgedrückt: von den 210 Personen, die ursprünglich zur Teilnahme an den Nachbefragungen bereit waren, konnten 154 (73% von 210) tatsächlich dreimal befragt werden (detaillierte Angaben zur Ausschöpfung finden sich in Tabelle 1).

Tabelle 1: Erwartete und realisierte Ausschöpfung der Test-Retest-Studie

		Erwartet	Realisiert
a	Haushalts-Adressen	420	420
b	Teilnehmer an der ersten Befragung (Hauptstudie)	294=70% von a	255=61% von a
c	Zur Teilnahme an Nachbefragungen bereit	235=80% von b	210=82% von b
d	Teilnehmer der zweiten Welle	200=85% von c	181=86% von c
e	Teilnehmer der dritten Welle	150=75% von d	154=85% von d

Ein zentrales Problem jeder Panel-Studie ist die sog. "Sterblichkeit" (Campbell/Stanley 1966), also das Wegfallen von Befragungspersonen über die Zeit. Da es in der Test-Retest-Studie zum ALLBUS 1984 in fast allen Fällen gelungen war, Kontakt zur Zielperson aufzunehmen, können die Gründe für den Ausfall von Befragungspersonen reproduziert werden (vgl. Tabelle 2).

Tabelle 2: Ausfallgründe in der Test-Retest-Studie

	Ausfall in der ...					
	2. Welle		3. Welle		Gesamt	
	N	%	N	%	N	%
a) Nichterreichbarkeit	4	13.8	4	14.8	8	14.3
Urlaub, Dienstreisen	6	20.7	8	29.6	14	25.0
b) Krankheit	4	13.8	6	22.2	10	17.9
c) Verweigerungen, Abbrüche	2	6.9	-	-	2	3.6
"Zu persönliche Fragen"	2	6.9	-	-	2	3.6
"Fragengleichheit"	5	17.2	-	-	5	8.9
Kein Interesse	4	13.8	3	11.1	7	12.5
Keine Zeit	2	6.9	4	14.8	6	10.7
Trotz Terminabsprache nicht anzutreffen	-	-	2	7.4	2	3.6
	29	100.0	27	99.9	56	100.1

Aus der Tabelle lassen sich drei Hauptarten von Ausfällen ablesen, nämlich a) Nichterreichbarkeit, b) befragungsunabhängige Ausfälle (Krankheit) und c) "reaktive" Ausfälle, die man als mehr oder minder massive Verweigerungen zu interpretieren hat.

Um zu vermeiden, daß die Befragungsergebnisse über die Zeit durch einen Wechsel des Interviewers zwischen den drei Wellen beeinflußt würden, sollten die Nachbefragungen von jeweils dem gleichen Interviewer durchgeführt werden, der auch die Befragung in der ersten Welle ausgeführt hatte. Dies konnte in 130 oder 84% der 154 in allen drei Wellen realisierten Interviews auch erreicht werden. In 23 Fällen waren zwei Interviewer an der Realisierung der drei Interviews beteiligt, in einem Fall drei Interviewer.

Die Interviews der Nachbefragungen wurden in der Zeit zwischen dem 15. April und 8. August 1984 durchgeführt. Ausgehend von den Interviews in der Haupterhebung (Feldzeit: 12. März - 30. Mai 1984) sollte jede Befragungsperson nach exakt vier Wochen zum ersten, nach exakt weiteren vier Wochen zum zweiten Mal nachbefragt werden.

Im Laufe der Feldzeit zeigte sich schnell, daß diese Vorgabe zu restriktiv war. Oft konnte - trotz prinzipieller Bereitschaft eines Befragten zur Teilnahme an weiteren Befragungen - kein Termin in der geplanten Erhebungswoche

realisiert werden. Um die angestrebte Ausschöpfungsquote erreichen zu können, wurden die Abstände zwischen den Befragungen frühzeitig neu definiert: Nicht mehr die vierte und die achte Woche nach der Befragung in der ersten Welle allein sollten zulässig sein, sondern - allerdings nur als Ausnahme von dieser Regel - eine Zeitspanne von der dritten bis zur fünften und von der siebten bis zur neunten Woche.

Alles in allem konnte aber auch diese weitergefaßte Vorgabe nicht völlig eingehalten werden. Der Abstand zwischen den Befragungen der ersten und zweiten Welle betrug zwar durchschnittlich 32 Tage, zwischen der zweiten und dritten Welle durchschnittlich 28 Tage, doch wurden 24% der Interviews in der zweiten Welle mehr als 5 Wochen nach der Befragung in der ersten Welle realisiert, 13% der Interviews in der dritten Welle mehr als fünf Wochen nach der Befragung in der zweiten Welle.

Nach Abschluß der Feldarbeiten wurde bei allen Befragten der dritten Panel-Welle eine telefonische Feldkontrolle durchgeführt. Ihre Ergebnisse führten dazu, alle 154 Fälle als korrekt realisiert einzustufen und in die Analysen mit einzubeziehen.

1.5 Qualität der Stichprobe - Vergleich der Panel-Stichprobe mit der Stichprobe des ALLBUS 1984

Um zu prüfen, inwieweit die Test-Retest-Stichprobe zumindest näherungsweise ein Abbild der Hauptstudie darstellt, wurden für eine große Anzahl ausgewählter Variablen die Häufigkeitsverteilungen der Hauptstudie (bezogen auf 3.004-154=2.850 Befragte) mit denen der Test-Retest-Teilstichprobe verglichen. Mit Hilfe des Chi^2-Tests wurde dann geprüft, ob die Verteilungen für die entsprechenden Variablen signifikant voneinander abwichen.

Bei dem Vergleich der Häufigkeitsverteilungen berücksichtigten wir vor allem die wichtigsten sozio-demographischen Variablen, die durch weitere objektive Merkmale wie die "frühere berufliche Stellung" und die "berufliche Stellung des Vaters" ergänzt wurden.

Neben diesen für einen Vergleich der ALLBUS-Hauptstudie und der Test-Retest-Stichprobe zentralen Merkmalen haben wir noch eine große Anzahl von Einstel-

lungsvariablen zu verschiedenen Themenbereichen nach der gleichen Vorgehensweise analysiert. Unser besonderes Interesse galt dabei den Fragen zu den Bereichen Wohlfahrtsstaat, Ungleichheit, Gastarbeiter und Politik.

Für die jeweiligen Itembatterien zu diesen Themenbereichen haben wir darüber hinaus die Kovarianzen berechnet. Wir haben dann auch hier die jeweiligen Kovarianzmatrizen der Befragten der Test-Retest-Studie mit den restlichen Befragten der Hauptstudie verglichen. Ein Vergleich der Kovarianzmatrizen stellt einen strengeren Test der Datenqualität der Test-Retest-Studie dar als der bloße Vergleich der Randverteilungen. Sind nämlich auch die bivariaten Verteilungen zwischen den Stichproben ähnlich, ist dies ein weiterer Beleg für die These, daß die Test-Retest-Stichprobe ein verkleinertes Spiegelbild der ALLBUS-Stichprobe ist, ein Ergebnis, das wegen möglicher Zufälligkeiten im Antwortverhalten aus dem Vergleich der Randverteilungen alleine nicht unbedingt zu erzielen ist.

1.5.1 Vergleich der Häufigkeitsverteilungen

Der Vergleich der Randverteilungen bei den demographischen Variablen[2] zeigt keine signifikanten Unterschiede zwischen der Test-Retest-Stichprobe und der Stichprobe der Hauptstudie. Einzige Ausnahme der analysierten demographischen Variablen ist das Alter des Befragten mit einem Chi_5^2-Wert von 24.43 bei einem Signifikanzniveau von $p < .001$. Allerdings hat dieser hohe Chi^2-Wert seine Ursache in der Berechnungsformel dieses Maßes.[3] Bei einer anderen Zusammenfassung der Alterskategorien treten keine signifikanten Unterschiede zwischen den Verteilungen auf.

Die - allerdings noch nicht signifikanten - Abweichungen beim Einkommen ($Chi_{10}^2 = 16.29$, $p = .092$) sind gleichermaßen auf die Kategorisierung zurückzuführen. Auch hier würde der Chi^2-Wert bei einer anderen Zusammenfassung der Kategorien (z.B. zu insgesamt 5 neuen Kategorien) ebenfalls dramatisch sinken, da sich dadurch die Abweichungen der Prozentwerte insbesondere der unteren Kategorien erheblich verringern würden.

Ein Überblick über die Ergebnisse des Vergleichs der verbleibenden demographischen Variablen ergibt sich aus Tabelle 3. Eine detailliertere Darstellung der Ergebnisse findet sich bei Zeifang (1987).

Tabelle 3: Vergleich ausgewählter demographischer Variablen zwischen ALLBUS-Hauptstudie und Test-Retest-Studie - Ergebnisse der Chi^2-Tests

Variable	Chi^2	df	p
Schulabschluß	2.692	4	0.611
Beruflicher Ausbildungsabschluß	4.025	7	0.777
Stellung im Erwerbsleben	5.982	8	0.649
Berufliche Stellung	9.130	5	0.104
Erste berufliche Stellung	6.237	7	0.512
Familienstand	3.883	4	0.422
Berufliche Stellung Vater	7.997	7	0.333
Konfession	9.233	5	0.100
Geschlecht	0.000	1	0.993

Zusammenfassend können wir festhalten, daß sich bei keiner sozio-demographischen Variablen signifikante Unterschiede zwischen den Häufigkeitsverteilungen der Stichprobe des ALLBUS 1984 und der Retest-Stichprobe zeigen, d.h. die Retest-Stichprobe bei den demographischen Variablen in der Tat ein verkleinertes Abbild der ALLBUS-Hauptstudie darstellt.

Neben den demographischen Variablen wurden - auf die gleiche Weise - ca. 50 Einstellungsitems überprüft. Der Vergleich dieser Items zwischen der Test-Retest-Stichprobe und den restlichen Befragten der Hauptstudie führte nur in einem Fall (Parteienthermometer für die DKP) zu einem signifikanten Unterschied[4] zwischen den beiden Stichproben.[5]

Obgleich gerade bei den Einstellungsvariablen in stärkerem Maße als bei den demographischen Variablen Unterschiede in den Häufigkeitsverteilungen zu erwarten waren, ist es doch außerordentlich bemerkenswert, daß auch bei diesen Variablen keine signifikanten Verzerrungen der Randverteilungen aufgetreten sind.

1.5.2 Vergleich der Kovarianzen ausgewählter Einstellungsitems

Obwohl bereits der Vergleich der univariaten Verteilungen der Einstellungsitems zwischen ALLBUS-Hauptstudie und Test-Retest-Studie ein durchgängig positives Ergebnis im Sinne unserer Erwartungen erbrachte, hielten wir es dennoch für erforderlich, auch bivariate Verteilungen (Kovarianzen) zwischen ausgewählten Items bzw. innerhalb bestimmter Itembatterien zu berechnen und zu vergleichen. Dabei haben wir uns auf die Itembatterien "Einstellungen zum Wohlfahrtsstaat", "Einstellungen zu sozialer Ungleichheit", "Einstellungen

zu Gastarbeitern" sowie "Einstellungen zu den politischen Parteien" (Parteienthermometer) beschränkt. Wenn die Test-Retest-Studie in der Tat ein Spiegelbild der ALLBUS-Hauptstudie sein soll, müßten auch die Kovarianzen der Items innerhalb dieser Einstellungsbatterien gleich oder zumindest sehr ähnlich sein.

Als Indikator für die Übereinstimmung zwischen den Kovarianzen dient auch hier der Chi^2-Test. Eine Übersicht über die Ergebnisse vermittelt Tabelle 4-:[6)]

Tabelle 4: Vergleich der Kovarianzmatrizen der Test-Retest-Studie und der Hauptstudie ALLBUS 1984 - Ergebnisse der Chi^2-Tests

	Chi^2	df	p
Einstellungen zum Wohlfahrtsstaat	37.05	36	.420
Einstellungen zu sozialer Ungleichheit	27.08	36	.858
Einstellungen zu Gastarbeitern	10.29	10	.416
Einstellungen zu politischen Parteien	40.64	28	.058

Aus der Tabelle ist zu ersehen, daß die Kovarianz-Strukturen der ALLBUS-Hauptstudie und der Test-Retest-Studie sehr ähnlich sind.[7)] Dies bestätigt die Ergebnisse der Vergleiche univariater Verteilungen und führt uns letztlich zu dem Schluß, daß die Test-Retest-Stichprobe tatsächlich ein verkleinertes Abbild der Stichprobe der ALLBUS-Hauptstudie darstellt.

2. Zur Stabilität von Umfragedaten - Vergleich ausgewählter Variablen

Die eigentlich zentrale substantielle Frage dieses Artikels beschäftigt sich mit der Stabilität von Umfragedaten. Bei der Verwendung solcher Daten in komplexen Analysemodellen und bei Zeitreihenanalysen ist Stabilität unabdingbare Voraussetzung: Erhebliche Antwortinstabilitäten bei der Beantwortung einzelner Fragen über die jeweiligen Wellen werden die Aufstellung von Modellen mit multiplen Indikatoren beeinflussen, wie auch manche ihrer Ergebnisse durch Rekurs auf den Dateninput eine plausible Erklärung erfahren werden.

Die Darstellung unserer Ergebnisse trennen wir nach demographischen Variablen und Einstellungsvariablen. Innerhalb dieser beiden Bereiche differenzieren wir noch einmal in kategoriale und intervallskalierte Variablen. Für

beide Variablengruppen werden Reliabiltäten und Stabilitäten berechnet. Die Reliabilitäten und Stabilitäten der intervallskalierten Variablen wurden nach Heise (1969) ermittelt. Da nach unserem Wissen bisher noch keine Methode existiert, um auch bei kategorialen Variablen Unzuverlässigkeit von wahrem Wandel zu trennen, können für diese Variablen keine entsprechenden Werte errechnet werden.[8] Um jedoch auch die Stabilitäten dieser Variablen zu prüfen, haben wir jeweils die stabilen Antworten über zwei bzw. drei Wellen ausgewiesen und die entsprechenden Assoziationsmaße (Cramer's V bzw. Tau B) berechnet.[9]

2.1 Demographische Variablen

Bei demographischen Variablen kommt der Antwortstabilität aus zwei Gründen besondere Bedeutung zu: Zum einen dienen sie in vielen Analysen von Umfragedaten allgemein als unabhängige Variablen, zum anderen sind die Ergebnisse unserer Analysen wichtig für die Weiterentwicklung standardisierter Instrumente zur Messung soziodemographischer Hintergrundsmerkmale in Umfragen.

a) Kategoriale Variablen

Bei Betrachtung der kategorialen Variablen über alle drei Wellen zeigen sich bei den prozentualen Antwortstabilitäten beim Geschlecht und beim Familienstand mit 99,4%[10] bzw. 98,1% Übereinstimmung äußerst befriedigende Stabilitätswerte.

Zufriedenstellende Antwortstabilitäten erhalten wir auch bei den Fragen nach dem allgemeinbildenden Schulabschluß (89,3%) und der Konfession des Befragten (89,0%) sowie dem Schulabschluß des Vaters (89,6%).

Als Ursache für die Veränderungen beim Schulabschluß des Befragten vermuten wir, daß nicht immer der höchste Schulabschluß angegeben wurde, daß ältere Befragte Schwierigkeiten hatten, ihren Schulabschluß in die Antwortkategorien einzuordnen (anderer Name des Abschlusses), oder daß von einigen Befragten der Schulabschluß - insbesondere in der ersten Welle - zu hoch angegeben wurde.

Bei den Befragten, die wechselnde Angaben über ihre Konfessionszugehörigkeit machen, nehmen wir an, daß Personen, die aus der Kirche ausgetreten sind bzw. konvertierten, in einer Welle die jetzige bzw. keine Konfession genannt und in einer anderen Welle ihre ursprüngliche Konfession angegeben haben.

Die auf den ersten Blick Überraschend hohen Werte beim Schulabschluß des Vaters resultieren daher, daß fast 80% der Väter einen Volks- bzw. Hauptschulabschluß haben und somit keine großen Klassifikationsleistungen vom Befragten gefordert waren.

Die Stabilitäten der weiteren ausgewählten sozio-demographischen Variablen sind weniger zufriedenstellend. Variablen, deren Antwortstabilitäten über alle drei Wellen über 70% liegen und deshalb von uns noch als akzeptabel bezeichnet werden, sind "derzeitige berufliche Erwerbstätigkeit" (81,2%), "überwiegender Lebensunterhalt" (78,4%), "derzeitige berufliche Stellung" (73,0%) und "beruflicher Ausbildungsabschluß" (72,0%). Warum sind diese Stabilitäten niedriger als eigentlich erwartet?

Da alle vier Items über relativ viele Antwortkategorien (zwischen 8 und 32 Kategorien) verfügen, ist eine perfekte Antwortübereinstimmung über alle drei Wellen von vornherein nicht zu erwarten. Daneben dürften bei den Fragen nach dem letzten beruflichen Ausbildungsabschluß, der derzeitigen Erwerbstätigkeit und dem überwiegenden Lebensunterhalt diejenigen Befragten Schwierigkeiten haben, sich immer in die gleiche Antwortkategorie einzuordnen, die mehrere Ausbildungsabschlüsse haben, verschiedene Erwerbstätigkeiten ausüben und/oder ihren Lebensunterhalt mit mehreren Einkommensarten bestreiten.

Trotz der relativ niedrigen Stabilitäten über alle drei Wellen sollte man allerdings nicht unberücksichtigt lassen, daß die prozentualen Übereinstimmungen beim Vergleich der einzelnen Wellen bei allen vier Items zwischen 75,0% und 91,6% liegen und somit als relativ hoch anzusehen sind. Schließlich sind auch die jeweiligen Assoziationsmaße für alle vier Variablen sehr hoch.

Variablen, deren Stabilitäten über alle drei Wellen als nicht akzeptabel angesehen werden müssen, sind "letzte berufliche Stellung" (55,7%), "berufliche Stellung des Vaters" (52,1%) und "erste berufliche Stellung" (42,0%).

Dies ist jedoch nicht überraschend, wenn wir uns vergegenwärtigen, daß bei diesen Fragen von den Interviewten jeweils eine beträchtliche Klassifikationsleistung gefordert wird: Die Befragten müssen ihre jeweilige berufliche Stellung bzw. die ihres Vaters einer der 32 (bzw. beim Vater sogar 37) vorgegebenen Antwortkategorien zuordnen.

Alle diese Ergebnisse bestätigen eine allgemeine Erklärung für die Antwortstabilität von Variablen: Je größer die Anzahl der Antwortkategorien ist, um so mehr nimmt die Antwortstabilität ab, "da aufgrund der geringen Bandbreite der Interpretation einzelner Antwortkategorien sich Antwortunsicherheiten in einer Wiederholungsbefragung leichter in der Wahl einer anderen Antwortkategorie niederschlagen können" (Koch 1985:30).

Betrachten wir im folgenden die vier Fragen nach den beruflichen Stellungen etwas genauer: Wie Koch (1985:67f.) zu Recht anmerkt, muß der Interviewte bei der Beantwortung dieser Fragen zwei Arten von Differenzierungen leisten: Zunächst muß er unter den rechtlich-institutionell definierten Stellungen im Beruf "seine" Stellung wiederfinden (z.B. Arbeiter, Angestellter). Im zweiten Schritt muß er dann eine Differenzierung im hierarchischen Niveau innerhalb des gewählten Oberbegriffs vornehmen (ungelernter Arbeiter, angelernter Arbeiter usw.).

Wie aus der nachstehenden Tabelle 5 hervorgeht, scheinen die Befragten mit der ersten Differenzierung (Einordnung in die Hauptkategorien) wenig Probleme zu haben. Sehen wir von der Frage nach der "ersten beruflichen Stellung" ab, so können die Antwortstabilitäten als sehr gut bezeichnet werden.

Tabelle 5: Antwortstabilitäten über alle drei Wellen für die Fragen nach den beruflichen Stellungen

	Nach 7 bzw. 8 Hauptgruppen differenziert			Nach 32 bzw. 37 Untergruppen differenziert			Prozentsatzdifferenz
	N	S	%	N	S	%	%
Derzeitige berufliche Stellung	63	62	98.4	63	46	73.0	25.4
Letzte berufliche Stellung	61	55	90.2	61	34	55.7	34.5
Erste berufliche Stellung	131	73	55.7	131	55	42.0	13.7
Berufliche Stellung des Vaters	140	113	80.7	140	73	52.1	28.6

N = Befragtenzahl, S = Anzahl stabiler Antworten,
% = Anteil stabiler Antworten von N

Offensichtlich stellt das eigentliche Problem für den Befragten die Einordnung "seiner" beruflichen Stellung in die entsprechende Subkategorie dar, d.h. wenn der Befragte nach 32 bzw. 37 Subkategorien differenzieren muß, sinken die Antwortstabilitäten über alle drei Wellen drastisch ab (bis zu 34,5%).

Neben dieser zu großen Anzahl von Antwortkategorien sind die verschiedenen Vorgaben in den Subkategorien teilweise den Befragten nicht geläufig (zu generelle Bezeichnungen), oder sie sind nicht trennscharf genug.[11)]

Als Folgerung aus diesen Ergebnissen schlagen wir für zukünftige Befragungen vor, diese Frage zweigeteilt zu erheben (Frage 1 nach der Hauptkategorie, Frage 2 nach der zugehörigen Subkategorie), und die Subkategorien präziser zu formulieren.

Sehr hohe Antwortstabilitäten zeigen sich bei einer weiteren Fragengruppe, nämlich den fünf Fragen nach den beruflichen Tätigkeiten. Bei keiner dieser "offen" gestellten Fragen (derzeitige berufliche Tätigkeiten von Selbständigen und Nicht-Selbständigen, erste und letzte berufliche Tätigkeit sowie die berufliche Tätigkeit des Vaters) fallen die Stabilitäten über alle drei Wellen unter 78%.

Wie Koch hierzu ausführt, entspricht die Aufforderung, eine Berufsbezeichnung unter Bezug auf die Tätigkeitsinhalte zu nennen, einer weitgehend im Alltag üblichen Konvention der Berufsklassifikation. Da die Nennung einer solchen konkreten Berufsbezeichnung für die Befragten die Wiederholung einer oft geübten Leistung darstellt (Koch 1985: 66), fallen die Antwortstabilitäten entsprechend hoch aus (zum Zusammenhang zwischen Fragen im Interview und den Konzepten, mit denen Personen ihre Erfahrungen codieren, vgl. Cannell/Kahn 1968:558).

Diese fünf Fragen ohne standardisierte Antwortvorgaben werden also von den Befragten sehr reliabel beantwortet. Die sich aus dieser Feststellung unmittelbar ergebende Frage, ob und falls ja in welchem Ausmaß "offene" Fragen zuverlässiger beantwortet werden als standardisierte Fragen zum gleichen Sachverhalt, läßt sich ebenfalls mit der Test-Retest-Studie untersuchen. Die Frage nach der Branche, in der der Befragte derzeit arbeitet, wurde sowohl

offen als auch mit standardisierten Antwortvorgaben (31 Antwortkategorien) vorgelegt.

Ausgehend von der Überlegung, daß die Zuordnung einer Branchenbeschreibung zu einem Wirtschaftszweig durch speziell qualifizierte Vercoder besser geleistet wird als durch die Befragten, müßte die Reliabilität der offenen Frage höher sein als die der geschlossenen. Diese Hypothese wird durch die Daten bestätigt: Während 93,7% aller Befragten über alle drei Wellen die offene Frage stabil beantworten, sind dies bei der geschlossenen Frage "nur" 83,9%.

Als erstes Zwischenresümee können wir festhalten, daß Fragen, die komplexe, aber für den Befragten unmittelbar alltagsrelevante Probleme zum Thema haben, möglicherweise besser in offener Form gestellt und von geschulten Vercodern klassifiziert werden sollten. Fragen mit vorgegebenen Antwortkategorien werden dann sehr reliabel beantwortet, wenn die Antwortvorgaben klar und eindeutig formuliert, gegeneinander klar abgegrenzt und die Informationen für die Befragten leicht "abrufbar" sind.

b) Intervallskalierte Variablen

Bei den intervallskalierten sozio-demographischen Variablen erhalten wir erwartungsgemäß hohe Antwortstabilitäten sowohl bei der Frage nach dem Alter wie auch bei der Kinderzahl der Befragten. Die Stabilitäten über alle drei Wellen erreichen bei beiden Fragen fast 100%.

Beim Vergleich der Antworten über die drei Wellen für die beiden Fragen nach dem Alter bei der ersten Heirat (für verheiratete bzw. verwitwete oder geschiedene Befragte) und bei den Angaben zur wöchentlichen Arbeitszeit können die Ergebnisse als noch befriedigend bezeichnet werden.[12)]

Auf den ersten Blick unbefriedigende Resultate ergeben sich bei den restlichen intervallskalierten Demographievariablen. Vor allem bei der für Sozialstrukturanalysen zentralen Variablen "Einkommen" machen nur 26,8% der Befragten über alle drei Wellen exakt dieselbe Angabe.

Allerdings sind diese Abweichungen nicht so dramatisch, wie es zunächst den Anschein hat: Wie wir den Daten entnehmen können, geben viele Befragte in einer Welle das genaue Einkommen an, z.B. DM 2.030,-, machen aber in der

nächsten Welle nur noch eine gerundete Angabe, in diesem Fall also DM 2.000,-. Dies wird auch evident bei einer Betrachtung der Korrelationskoeffizienten: So zeigen sich in allen drei Beziehungen zwischen den Wellen Korrelationskoeffizienten von über 0.900, d.h. die Zusammenhänge der Einkommensangaben sind zwischen den einzelnen Wellen sehr hoch.

Darüber hinaus weisen die nach Heise (1969) berechneten Schätzwerte für die Reliabilitäten und die Stabilitäten sämtlicher intervallskalierter Demographievariablen äußerst befriedigende Ergebnisse auf. Für alle Variablen betragen die Reliabilitäten mindestens .968; die Stabilitäten zwischen den Wellen liegen bei fast allen Variablen zwischen .900 und 1.000.

Abschließend sei noch auf ein äußerst interessantes Ergebnis verwiesen, das im nächsten Kapitel eingehender diskutiert werden soll: Bei fast allen demographischen Variablen sind die Antwortstabilitäten zwischen der zweiten und der dritten Welle höher als diejenigen zwischen Welle 1 und 2 bzw. 1 und 3.

2.2 Einstellungsvariablen

Vergleichen wir die Antwortstabilitäten der Einstellungsfragen mit denen der demographischen Variablen, so zeigen sich erwartungsgemäß bei fast allen Einstellungsitems erheblich niedrigere Antwortstabilitäten sowohl im Vergleich von je zwei wie auch über alle drei Wellen.

Lediglich bei der Frage nach der subjektiven Schichteinstufung sowie bei den Fragen nach ihrem vergangenen bzw. zukünftigen Wahlverhalten scheinen die Befragten festgefügte Meinungen zu haben bzw. sich gut an ihre letzte Wahlentscheidung zu erinnern: Rund 80% aller Interviewten gaben nämlich bei allen drei Befragungen jeweils die gleiche Antwort.

a) Kategoriale Variablen

Bei den kategorial skalierten Fragen zum Wohlfahrtsstaat zeigen sich nur bei drei Items (Item D: "Staatliche Fürsorgepflicht", Item F: "Gutes Leben in der BRD" und Item G: "Gerechte Verteilung") Antwortstabilitäten von rund 50% über alle drei Wellen. Zu diesen Items, die eher generelle Einstellungen zum Wohlfahrtsstaat thematisieren, existiert bei den Befragten offensichtlich auch ein relativ festgefügtes Meinungsbild.

Neben dieser eher inhaltlichen Erklärung dürfte jedoch auch noch ein fragetechnischer Faktor für die höheren Stabilitäten gegenüber den anderen fünf Items[13] verantwortlich sein: Alle drei Items weisen jeweils nur einen Stimulus auf, während dies bei den restlichen Items meist nicht der Fall ist.[14] So könnte beispielsweise eine Befragungsperson bei Item A in der ersten Welle als Stimulus den ersten Satz des Items ("Jeder muß für sich selbst sorgen"), bei der nächsten Welle den zweiten Satz mit Betonung auf politischem Kampf und bei der letzten Welle diesen zweiten Satz mit Betonung auf gewerkschaftlichem Kampf beantwortet haben.[15]

Während bei einigen Items zum Wohlfahrtsstaat noch Stabilitäten von ca. 50% aufgetreten sind, finden sich bei den Ungleichheitsitems keine Stabilitäten in dieser Höhe. Bei einem einzigen Item dieser Batterie geben noch über 40% der Befragten dreimal die gleiche Antwort, die anderen Items werden nur noch von ca. jedem dritten Befragten über alle drei Wellen stabil beantwortet. Item A fällt in der Stabilität völlig ab und weist mit 22.5% einen äußerst niedrigen Wert auf. Wir führen dieses Ergebnis vor allem auf den doppelten Stimulus, der in diesem Item enthalten ist, zurück.[16]

Neben dem Problem doppelter Stimuli dürfte für die relativ niedrigen Stabilitäten der Wohlfahrts-, aber in noch stärkerem Maße der Ungleichheitsitems die mangelnde Trennschärfe der Antwortkategorien verantwortlich sein.[17]

Analysieren wir die Häufigkeitsverteilungen pro Welle, so sehen wir, daß die meisten Befragten eine der beiden mittleren Antwortkategorien angeben (ausweichen?) und sich die Instabilitäten vor allem aufgrund eines Pendelns zwischen diesen beiden Antwortkategorien ergeben.

Die starke Konzentration auf eine der beiden mittleren Antwortkategorien dürfte im übrigen auch darauf zurückzuführen sein, daß sich einige Interviewte durch die Fragen zu den Themen Wohlfahrtsstaat und Ungleichheit intellektuell überfordert fühlten oder sich zu diesen Fragen noch keine Meinung gebildet hatten. Anstatt nun mit offener Meinungslosigkeit zu reagieren (also mit "weiß nicht" zu antworten), versuchten sie wohl, eine inhaltliche Antwort zu geben; in der Regel wichen sie auf die mittleren Antwortkategorien aus. Eher zufällig entschieden sie sich dann für die eine oder die andere der beiden mittleren Antwortkategorien.

Gerade die Frage, ob sich eine Befragungsperson bereits mit einem Thema auseinandergesetzt hat bzw. ob ein Themengebiet für den Befragten eine bestimmte Relevanz hat, wirkt sich erheblich auf die Antwortstabilität und auch die Reliabilität aus. So weist Converse (1970) darauf hin, daß die unterschiedliche Zentralität, die ein Erhebungsobjekt für den Befragten hat, für die Höhe der Fluktuation seiner Antworten in Panelstudien entscheidend ist. Leider können wir diese These mit unserem Datenmaterial nicht weiter verfolgen.

Betrachten wir noch die Ergebnisse der Fragen zu politischen Wertorientierungen (Inglehart-Index): Die höchsten Stabilitätswerte über alle drei Wellen erhalten wir beim wichtigsten und beim unwichtigsten Ziel, d.h. die Befragten scheinen diese beiden Ziele noch - relativ gesehen - eindeutig festlegen zu können, während das zweit- und das drittwichtigste Ziel eher zufällig bestimmt werden. Dies wirkt sich insofern dramatisch aus, als das zweitwichtigste Ziel für die Zuordnung der Befragten zu den Inglehartschen Wertetypen von zentraler Bedeutung ist. Der in vielen empirischen Arbeiten verwendete Index ist also nach den Daten dieser Studie auf Individualebene über die Zeit ein instabiles Instrument.

b) Intervallskalierte Variablen

Auf den ersten Blick weisen die vier intervallskalierten Gastarbeiter-Items keine befriedigenden Antwortstabilitäten auf. Nur jeweils jeder vierte Befragte nennt in allen drei Wellen denselben Skalenwert für ein bestimmtes Item. Dieses Ergebnis ist jedoch nicht überraschend, wenn wir uns vergegenwärtigen, daß dem Befragten jeweils eine Skala mit 7 Ausprägungen vorgelegt wurde. Hier gilt sicherlich, was Stadtler (1980:8) ebenfalls im Zusammenhang mit nicht-verbalen Siebener-Skalen feststellt, nämlich "daß Skalen mit relativ vielen Kategorien die Personen hinsichtlich ihres Differenzierungsvermögens oder ihrer Differenzierungswilligkeit überfordern".

Wesentlich aussagekräftiger als die Antwortstabilitäten von Randverteilungen sind für die Beurteilung der Items sicherlich die aus den Korrelationskoeffizienten errechneten Reliabilitäten und Stabilitäten der Items. Während die Reliabilitäten mit Werten zwischen .755 und .956 sehr hoch sind, können die Stabilitäten zwischen den Wellen bis auf wenige Ausnahmen[18)] als bestenfalls zufriedenstellend bezeichnet werden.

Wenden wir uns abschließend den Antwortstabilitäten der Parteienthermometer zu. Von den beiden extremen Parteien NPD und DKP abgesehen bewegen sich die Antwortstabilitäten über alle drei Wellen nur zwischen 18.3 und 27.5%. Allerdings sollten wir bei diesen Werten beachten, daß wir es hier mit Antwortskalen zu tun haben, die dem Befragten 11 (!) Möglichkeiten der Einordnung bieten, d.h. wir müssen von vornherein niedrige Antwortstabilitäten erwarten. Die sehr hohen Stabilitäten für die beiden radikalen Parteien NPD und DKP scheinen zwar diese Aussage zu relativieren, sind jedoch darauf zurückzuführen, daß die Mehrzahl der Befragten diese Parteien in allen drei Wellen extrem schlecht (meist mit -5) bewertet hat.

Wesentlich bedeutsamer sind auch bei der Analyse der Ergebnisse zum Parteienthermometer die Reliabilitäten und Stabilitäten der Items. Sowohl die Reliabilitäten mit Werten zwischen .700 und .856 als auch die Stabilitäten mit Werten zwischen .900 und 1.000 sind für die Items beeindruckend hoch. Lediglich die Stabilitäten für die SPD insgesamt und für die F.D.P. zwischen Welle 1 und 3 (Stabilitäten zwischen .664 und .817) müssen als nicht befriedigend eingestuft werden. Offensichtlich haben aber die Befragten der Test-Retest-Studie ein relativ festgefügtes Meinungsbild von den sechs Parteien, das in den hohen Stabilitäten zum Ausdruck kommt.

Abschließend kommen wir zu einem der zentralen Ergebnisse der Test-Retest-Studie, auf das bereits im vorhergehenden Kapitel kurz hingewiesen wurde. Wir haben unsere bisherigen Analysen vor allem auf die Höhe der Antwortstabilitäten über alle drei Wellen konzentriert. Vergleichen wir dagegen die Stabilitäten zwischen den jeweiligen Befragungswellen, so können wir eine interessante Regelmäßigkeit konstatieren: Bei fast allen demographischen und Einstellungsvariablen sind die Antwortstabilitäten zwischen der zweiten und der dritten Welle höher als zwischen der ersten und der zweiten bzw. der ersten und der dritten Welle.[19)]

Als Ursache für dieses Ergebnis vermuten wir, daß sich die Befragten erst beim zweiten Interview bewußt waren, daß sie noch ein drittes Mal befragt würden[20)] und sie deshalb die beiden letzten Interviews ernsthafter durchgeführt haben, d.h. valider als in der Hauptstudie geantwortet haben.

Darüber hinaus ist es auch möglich, daß sich die Befragten nach dem ersten Interview mit den Themen der Umfrage auseinandergesetzt haben bzw. "unbewußt" Informationen zu den angesprochenen Themen gesammelt haben (vgl. dazu auch Jagodzinski/Kühnel/Schmidt 1987).

Welche von diesen beiden Erklärungen zutrifft, können wir anhand des vorliegenden Datenmaterials nicht entscheiden. Hierzu wären weitere Untersuchungen erforderlich.

3. Zusammenfassung und Bewertung

Neben der Beschreibung der Test-Retest-Studie zum ALLBUS 1984 setzten wir uns in diesem Artikel mit zwei Fragen auseinander.

Im ersten Teil sind wir der Frage nachgegangen, inwieweit die Test-Retest-Stichprobe repräsentativ für die ALLBUS-Haupterhebung ist. Vergleiche der Häufigkeitsverteilungen zentraler Variablen wie auch der Vergleich der Kovarianzmatrizen mehrerer Einstellungsfragen zeigten, daß die Test-Retest-Studie in der Tat ein verkleinertes Abbild der ALLBUS-Hauptstudie darstellt.

Im zweiten Teil diskutierten wir die Antwortstabilitäten ausgewählter Variablen über alle drei Wellen. Obwohl keine der demographischen Variablen eine Antwortstabilität von 100% über alle drei Wellen erreichte, waren die Stabilitäten von allen für die Analyse von Umfragedaten zentralen soziodemographischen Variablen außerordentlich hoch. Die Frage, wie stabil Umfragedaten eigentlich sind, findet, zumindest soweit es die demographischen Variablen angeht, an den Daten der Test-Retest-Studie eine ermutigende Antwort.

Im Vergleich zu diesen Variablen sind die Antwortstabilitäten über alle drei Wellen bei den Einstellungsvariablen erwartungsgemäß niedriger. Dies ist aber offensichtlich sehr häufig ein Resultat der Operationalisierung der Variablen.

Für die Höhe der Antwortstabilitäten scheinen vor allem fragetechnische Aspekte ausschlaggebend zu sein. So hat die Zahl der Antwortkategorien bzw. die Spannweite der Skalen einen erheblichen Einfluß auf die Stabilitäten, d.h. je geringer die Anzahl der Antwortkategorien ist bzw. je kürzer die Skalen sind, um so stabiler sind in der Regel die Antworten und vice versa.

Erhebliche Auswirkungen auf die Antwortstabilitäten zeigen sich darüber hinaus, wenn Fragen nicht eindeutig formuliert sind und mehr als einen Stimulus enthalten, oder die Antwortkategorien nicht ausreichend trennscharf sind.

Ein weiteres zentrales Ergebnis ist, daß die Antwortstabilitäten zwischen der zweiten und dritten Welle bei fast allen Variablen höher sind als diejenigen zwischen Welle 1 und 2 bzw. Welle 1 und 3. Wir können nur vermuten, daß bei den Befragten zwischen der ersten und zweiten Welle "verschiedene Sensibilisierungs-, Motivations- und Lernprozesse" (vgl. Koch 1985:61) abgelaufen sind, die sich bereits auf das Antwortverhalten in der zweiten Welle ausgewirkt und letztlich ihren Niederschlag in den höheren Antwortstabilitäten zwischen Welle 2 und 3 gefunden haben.

Dieser Artikel wurde von den früheren ALLBUS-Projektmitarbeitern Rolf Porst und Klaus Zeifang verfaßt, die auch die Test-Retest-Studie im Rahmen des ALLBUS-Projektes bearbeitet haben. Abschnitt 2.1 wurde unter Mitarbeit von Achim Koch erstellt, der Mitarbeiter des ALLBUS-Projekts ist.

Anmerkungen

1) Ein "Stichprobennetz" stellt eine systematische Unterstichprobe aus den ca. 50.000 Stimmbezirken der Bundesrepublik und West-Berlins bei Wahlen zum Deutschen Bundestag bzw. zum Berliner Abgeordnetenhaus dar (zur Stichprobenziehung vgl. Kirschner 1984). Jedes Stichprobennetz besteht aus 210 sample points (= Stimmbezirke bzw. synthetische Stimmbezirke).
2) Die ausführliche Dokumentation des Vergleichs der Randverteilungen sowie die Kovarianzmatrizen finden sich bei Zeifang (1987).
3) In der letzten Alterskategorie "89 und mehr Jahre" ist in der Hauptstudie nur eine Person enthalten, so daß die erwartete Zellenhäufigkeit mit 0.05 äußerst niedrig ist. Da die erwartete Zellenhäufigkeit u.a. im Nenner der Berechnungsformel für Chi^2 steht, erhöht sie den Chi^2-Wert für diese Kategorie erheblich (um 16.457 Punkte), d.h. ohne diese Alterskategorie würde der Chi^2-Wert mit 7.969 bei vier Freiheitsgraden keine Unterschiede der beiden Verteilungen signalisieren.
4) Die Ursache für diese Verzerrung ist wieder in der Berechnungsformel des Chi^2-Werts zu sehen: Vor allem die drei Zellen mit erwarteten Zellenhäufigkeiten, die kleiner als 1 sind, tragen zu einer exorbitanten Erhöhung des Chi^2-Werts bei. Ein visueller Vergleich der beiden Verteilungen zeigt uns jedoch, daß die Prozentwerte der beiden Stichproben ungefähr miteinander übereinstimmen.
5) Eine detaillierte Darstellung der Ergebnisse findet sich ebenfalls bei Zeifang (1987).
6) Detaillierte Informationen über Kovarianzen, Mittelwerte und Standardabweichungen sind bei Zeifang (1987) zu finden.
7) Lediglich beim Parteienthermometer zeigen sich kleinere Abweichungen, die jedoch noch nicht signifikant sind.
8) Mit den Auswirkungen mangelnder Reliabilität bei dichotomen Variablen beschäftigt sich Schwartz (1985); vgl. auch Koch (1985:16ff.).

9) Eine alternative Vorgehensweise wäre durch die Berechnungen von entsprechend angepaßten log-linearen Modellen möglich gewesen (vgl. Bishop/Tienberg/Holland 1975). Da jedoch die wesentlichen Informationen über die Itemstabilität auch durch die von uns gewählte einfachere Methode erhältlich sind, haben wir uns für diese Vorgehensweise entschieden.
10) Die Abweichungen der Geschlechtsangabe in der dritten Welle bei einer Person sind wahrscheinlich auf einen Interviewerfehler zurückzuführen.
11) Dies dürfte insbesondere auf die fünf Angestelltenkategorien zutreffen.
12) Die Prozentwerte für diese beiden Fragen betragen 69.4% bzw. 83.7%.
13) Die Antwortstabilitäten über alle drei Wellen liegen bei diesen Items zwischen 30.6% und 38.5%.
14) Diese Mehrdeutigkeit war in Kauf genommen worden, um diese Itembatterie aus einer älteren Studie exakt replizieren zu können.
15) Das Item hieß: "In unserer Gesellschaft muß jeder für sich schauen, daß er auf einen grünen Zweig kommt. Es hilft nicht viel, sich mit anderen zusammenzuschließen, um politisch oder gewerkschaftlich für seine Sache zu kämpfen".
16) Auch diese Itembatterie wurde aus einer älteren Studie exakt repliziert. Item A hieß: "In der Bundesrepublik bestehen noch die alten Gegensätze zwischen Besitzenden und Arbeitenden. Die persönliche Stellung hängt davon ab, ob man zu der oberen oder unteren Klasse gehört."
17) Die Antwortkategorien hießen: "stimme voll zu", "stimme eher zu", "stimme eher nicht zu", "stimme überhaupt nicht zu".
18) Die Antwortstabilitäten schwanken zwischen den Wellen von .591 bis .861. Nur für drei Items liegen sie über .900 zwischen Welle 2 und Welle 3.
19) Die Stabilitäten zwischen Welle 1 und 2 bzw. 1 und 3 sind häufig sehr ähnlich, wobei bei einigen Variablen die Stabilität zwischen Welle 1 und 2, bei anderen die zwischen Welle 1 und 3 höher ist.
20) Aus dem Text der Einwilligungserklärung ging hervor, daß die Befragten noch zweimal befragt werden könnten.

Literatur

Arminger, G., 1976: Anlage und Auswertung von Paneluntersuchungen. S. 134-235 in K. Holm (Hrsg.), Die Befragung, Band 4. München: Francke.
Bishop, Y.M.M./Tienberg, S.E./Holland, P.W., 1975: Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.
Campbell, D.T./Stanley, J.C., 1966: Experimental and Quasi-Experimental Designs for Research. Chicago, Ill.: Rand McNally.
Cannell, Ch./Kahn, R.L., 1968(2): Interviewing. S. 526-595 in G. Lindzey/E. Aronson (Hrsg.), The Handbook of Social Psychology, Bd. II. Reading, Mass.: Addison-Wesley.
Carmines, E.G./Zeller, R.A., 1979: Reliability and Validity Assessment. Beverly Hills: Sage.
Converse, Ph.E., 1970: Attitudes and non-attitudes: Continuation of a dialogue. S. 168-189 in E.R. Tufte (Hrsg.), Quantitative Analysis of Social Problems. Reading, Mass.: Addison-Wesley.
Heise, D.R., 1969: Separating reliability and stability in test-retest-correlation. American Sociological Review 34:93-101.
Heise, D.R./Bohrnstedt, G.W., 1970: Validity, invalidity and reliability. S. 104-129 in E.F. Borgatta/G.W. Bohrnstedt (Hrsg.), Sociological Methodology 1970. San Francisco: Jossey Bass.
Jagodzinski, W./Kühnel, S./Schmidt, P., 1987: Is there a "Socratic Effect" in non-experimental panel studies? Consistency of an attitude towards guestworkers. Sociological Methods and Research, Heft 2.

Jöreskog, K.G./Sörbom, D., 1977: Statistical models and methods for analysis of longitudinal data. S. 285-325 in D.J. Aigner/A.S. Goldberger (Hrsg.), Latent Variables in Socioeconomic Models. Amsterdam: North Holland.

Kessler, R.C./Greenberg, D.F., 1981: Linear Panel Analysis. Models of Quantitative Change. New York/London: Academic Press.

Kirschner, H.-P., 1984: ALLBUS 1980: Stichprobenplan und Gewichtung. S. 114-182 in K.-U. Mayer/P. Schmidt (Hrsg.), Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. Beiträge zu methodischen Problemen des ALLBUS 1980. ZUMA-Monographien Sozialwissenschaftliche Methoden, Band 5. Frankfurt/New York: Campus.

Koch, A., 1985: Wie zuverlässig lassen sich Berufs- und Bildungsvariablen messen? Diplomarbeit. Universität Mannheim.

Lord, F.M./Novick, M.R., 1968: Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley.

Porst, R./Schmidt, P., 1982: Analyse ausgewählter Meßinstrumente des ALLBUS 1980. Mannheim: unveröffentlicht.

Schwartz, J.E., 1985: The neglected problem of measurement error in categorical data. Sociological Methods and Research 13:435-466.

Stadtler, K., 1980: Die Auswirkungen unterschiedlicher Rating-Skalen auf das Urteilsverhalten von Befragten. Erste Ergebnisse. München: Infratest.

Wegener, B., 1983: Wer skaliert? Die Meßfehler-Testtheorie und die Frage nach dem Akteur. Theoretische Einleitung zum ZUMA-Handbuch Sozialwissenschaftliche Skalen. Mannheim: ZUMA und Bonn: IZ.

Wiley, D.E./Wiley, J.A., 1970: The estimation of measurement error in panel data. American Sociological Review 35:112-117.

Zeifang, K., 1987: Die Test-Retest-Studie zum ALLBUS 1984 - Tabellenband. ZUMA-Arbeitsbericht Nr. 87/01. Mannheim: ZUMA.